Supplementary Materials for

Spatially resolved, highly multiplexed RNA profiling in single cells

Kok Hao Chen, Alistair N. Boettiger, Jeffrey R. Moffitt, Siyuan Wang, and Xiaowei Zhuang.

correspondence to: zhuang@chemistry.harvard.edu

This PDF file includes:

Supplementary Figures

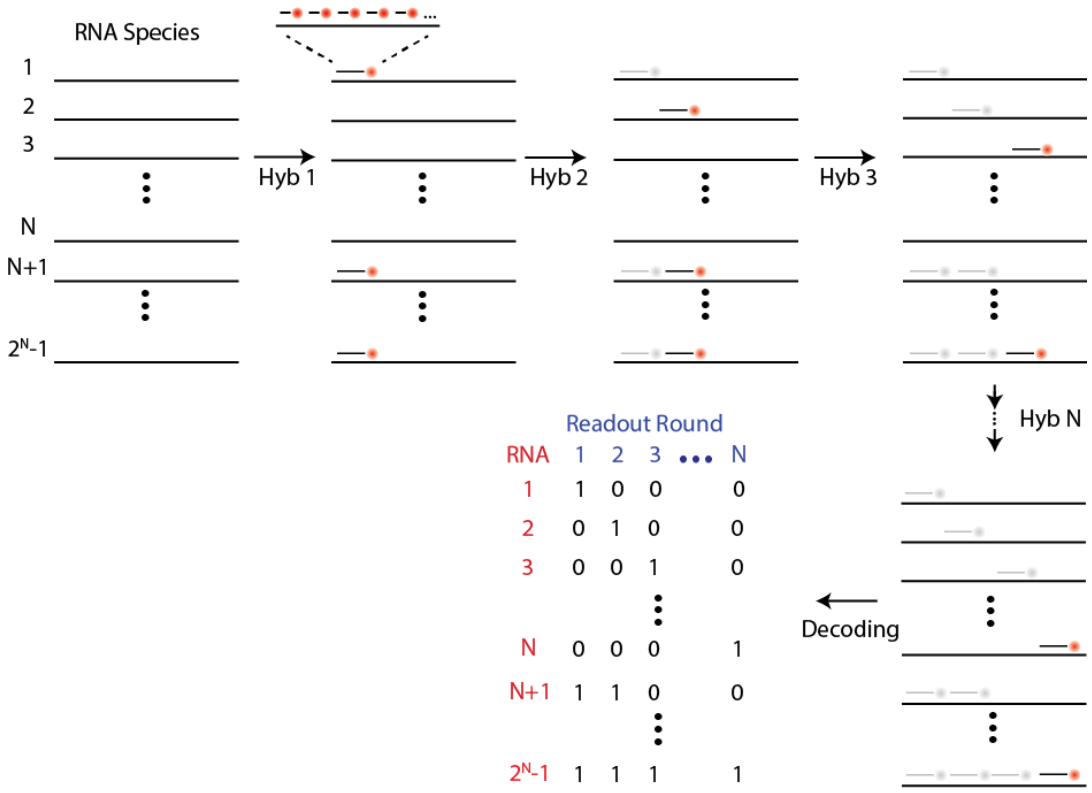

Fig. S1.

Schematic description of a combinatorial labeling approach based on a simple binary code. In a conceptually simple labeling approach, 2^N*-1* different RNA species can be uniquely encoded with all *N*-bit binary words (excluding the word with all '0's). In each hybridization round, FISH probes that are targeted to all RNA species that have a '1' in the corresponding bit are included. To increase the ability to discriminate RNA spots from background, each RNA is addressed with multiple FISH probes per hybridization round. Signal from the bound probes is extinguished before the next round of hybridization. This process continues for all *N* hybridization rounds (hyb 1, hyb 2, …), and all 2^N*-1* RNA species can be identified by the unique on-off pattern of fluorescence signals in each hybridization round.

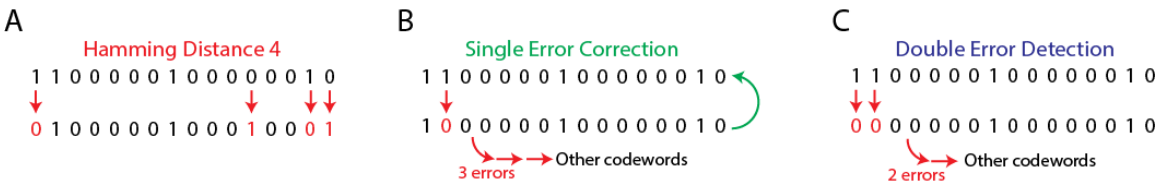

Fig. S2

Schematic descriptions of Hamming distance and its use in the identification and correction of errors. (**A**) Schematic representation of a Hamming distance of 4. (**B-C**) Schematic showing the ability of an encoding scheme with Hamming distance 4 to correct single-bit errors (B) or detect but not correct double-bit errors (C). Red arrows highlight bits at which the indicated words differ. Two code words are separated by a Hamming distance of 4 if one of the words has to flip four bits from '1' to '0' or '0' to '1' to convert into the other word. Single-bit error correction is possible because if a measured word differs from a legitimate code word by only one bit, it is most likely an error that arises from misreading this code word, since the code words of all the other RNA species will differ from the measured word by at least three bits. In this case, we can correct the measured word to the code word that differs by only one bit. If a measured word differs from a legitimate code word by two bits, this measured word can still be identified as an error, but correction is no longer possible since more than one legitimate code word differs from this measured word by two bits.

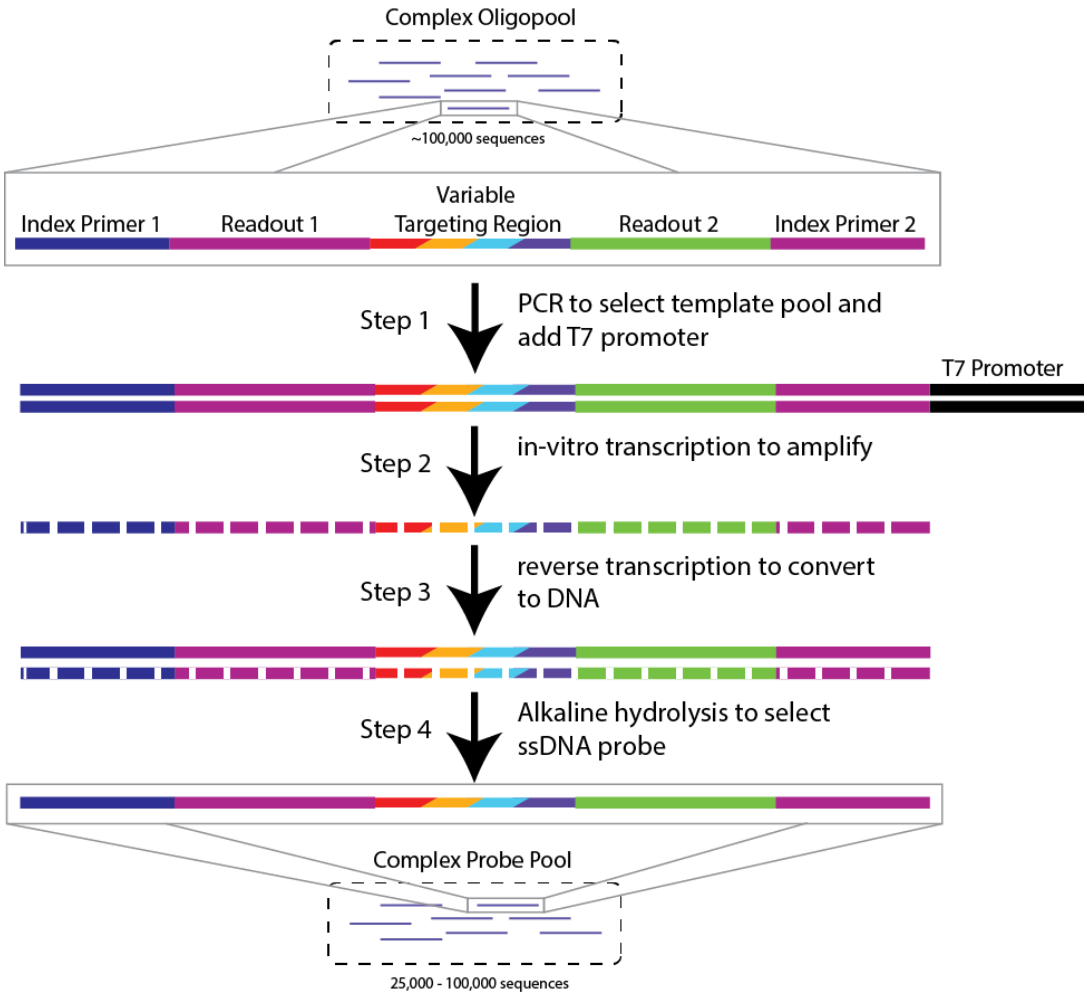

Fig. S3

Production of the library of encoding probes. An array-synthesized complex oligopool, containing ~100k sequences, is used as a template for the enzymatic amplification of the encoding probes for different experiments. Each template sequence in the oligopool contains a central target region that can bind to a cellular RNA, two flanking readout sequences, and two flanking index primers. In the first step, the required template molecules for a specific experiment are selected and amplified with an indexed PCR reaction. To allow amplification via *in vitro* transcription, a T7 promotor is added to the PCR products during this step. In the second step, RNA is amplified from these template molecules via *in vitro* transcription. In the third step, this RNA is reverse transcribed back into DNA. In the final step, the template RNA is removed via alkaline hydrolysis, leaving only the desired ssDNA probes. This protocol produces ~2 nmol of complex pools of encoding probes containing ~20,000 different sequences for the 140-gene experiments or ~100,000 different sequences for the 1001-gene experiments. This protocol is similar to a recently reported method but has achieved a substantially higher yield (*20*).

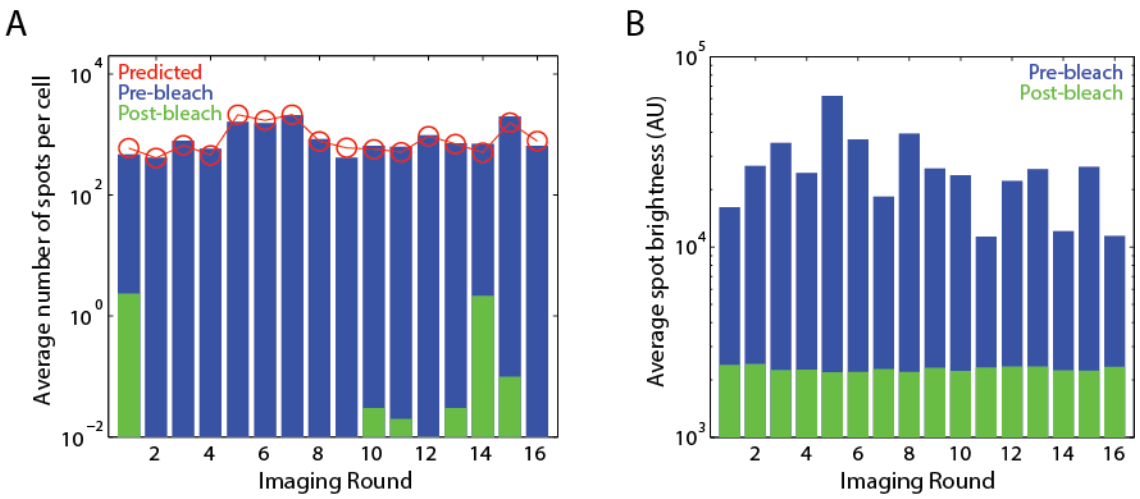

Fig. S4

The number and average brightness of the fluorescent spots detected in the 16 rounds of hybridization before and after photobleaching. (**A**) The number of fluorescent spots observed per cell before (blue) and after (green) photobleaching as a function of hybridization round averaged across all measurements with the first 16-bit MHD4 code. Photobleaching reduces the number of fluorescent spots by two or more orders of magnitude. Hybridization rounds without green bars represent rounds in which no molecules were observed after bleaching. Also depicted is the expected change in the number of fluorescent spots from round to round (red circles) predicted based on the relative abundances of the RNA species targeted in each hybridization round derived from bulk RNA sequencing. The average discrepancy between the observed and predicted number of spots for each hybridization is only 15% of the mean number of spots. This discrepancy does not systematically increase with the number of hybridization rounds. (**B**) The average brightness of the identified fluorescent spots in each hybridization round averaged across all measurements with the first 16-bit MHD4 code both before (blue) and after (green) photobleaching. Brightness varies by 40% (standard deviation) across different hybridization rounds. The variation pattern is reproducible between experiments with the same code, likely due to differences in the binding efficiency of the readout probes to the different readout sequences. There is a small systematic trend of decrease in the brightness with increasing hybridization rounds, which is on average 4% per round. Photobleaching extinguishes fluorescence to a level similar to that of the autofluorescence of the cell.

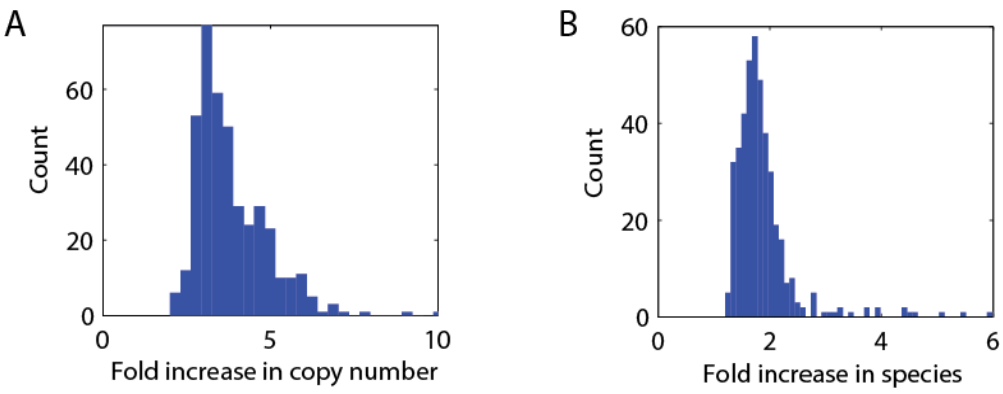

Fig. S5

Error correction substantially increases the numbers of RNA molecules and RNA species detected in individual cells. (**A**) Histogram of the ratio of the total number of molecules detected per cell with error correction to the number measured without error correction. (**B**) The histogram of the total number of RNA species detected in each cell with error correction to that without error correction. Both ratios are determined for ~200 cells and the histograms are constructed from these ratios.

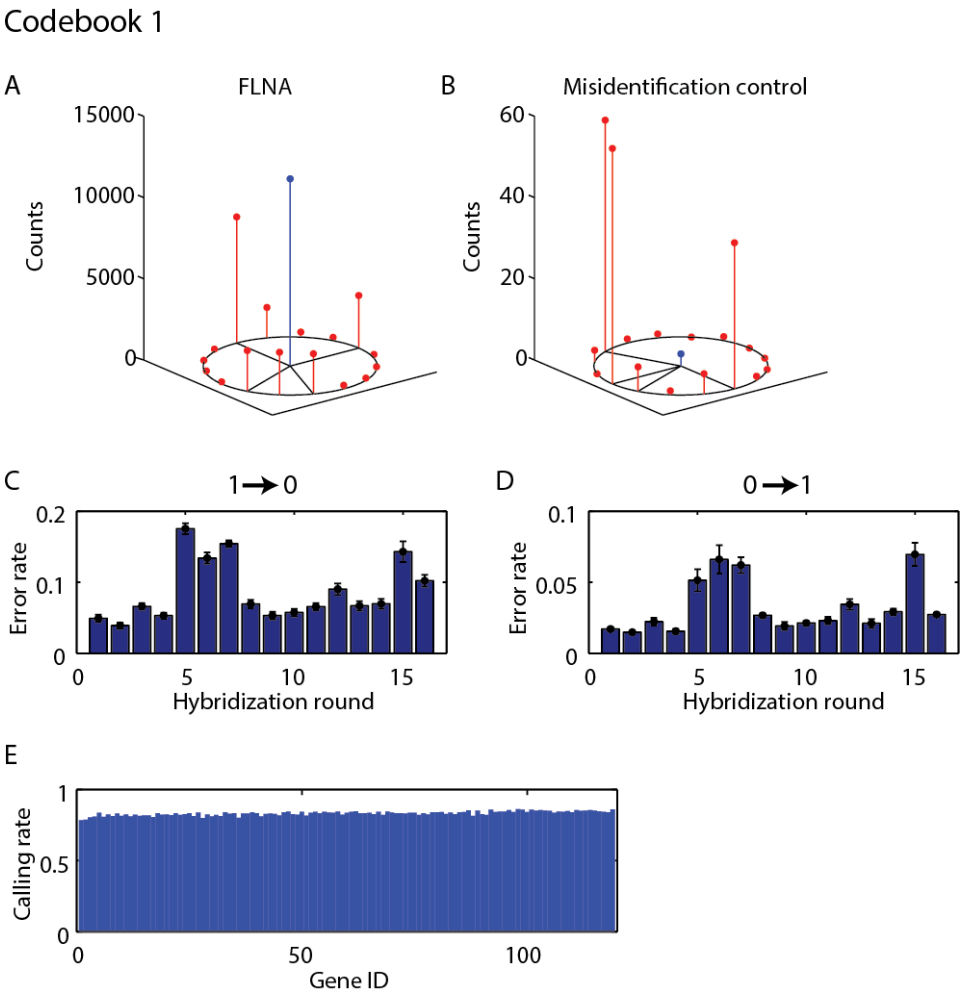

Fig. S6

Characterization of the misidentification and calling rates of RNA species for the 140-gene experiments using a specific 16-bit MHD4 code. (**A**) The number of measured words exactly matching the code word corresponding to FLNA, represented by the blue bar in the center of the circle, and the number of measured words with one-bit error compared to the code word of FLNA, represented by the 16 red bars on the circle. (**B**) The same as (A) but for a code word that was not assigned to any RNA — a misidentification control word. The solid lines connect the exact match to one-bit error words that are generated by 1→0 errors. Based on the observation that the ratio of the number of exact matches to the number of error-correctable matches for a real RNA-encoding word was typically substantially higher than the same ratios observed for the misidentification controls, we defined this ratio as a confidence ratio for RNA identification. The confidence ratio measured for all 130 RNA species (blue) and 10 misidentification control words not assigned to any RNA (red) using this 16-bit MHD4 code is show in Fig. 2F. (**C**, **D**) The average error rates for the 1→0 error (C) and 0→1 error (D) for each hybridization round. (**E**) The calling rate for each RNA species estimated from the 1→0 and 0→1 error rates. Genes are sorted from left to right based on the measured abundance, which spans three orders of magnitude. The calling rates are largely independent of the abundance of the gene.

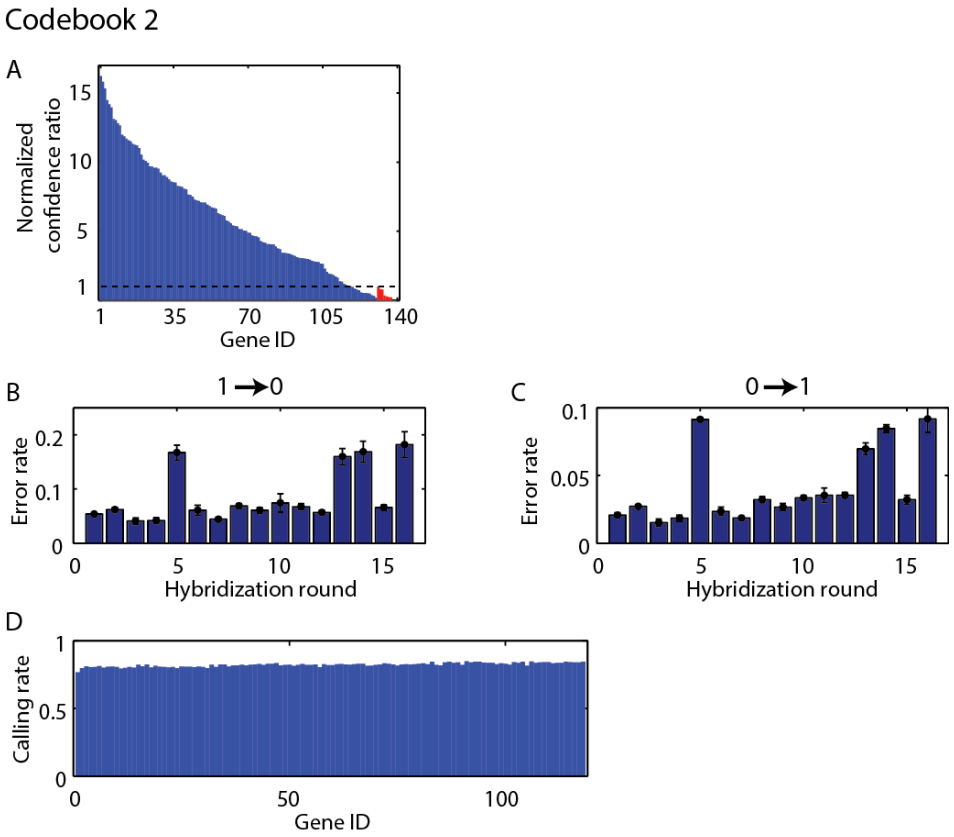

Fig. S7

Characterization of the misidentification and calling rates for a second 16-bit MHD4 code. In this second encoding scheme, we shuffled the 140 code words among different RNA species and changed the encoding probe sequences. (**A**) The normalized confidence ratio measured for the 130 RNA species (blue) and the 10 misidentification control words not assigned to any RNA (red). The normalized confidence ratio is determined the same way as in Fig. 2F. (**B**, **C**) The average error rates determined for the 1→0 error (B) and 0→1 error (C) for each hybridization round. (**D**) The calling rate determined for each RNA species estimated from the 1→0 and 0→1 error rates. Genes are sorted from left to right based on the measured abundance.

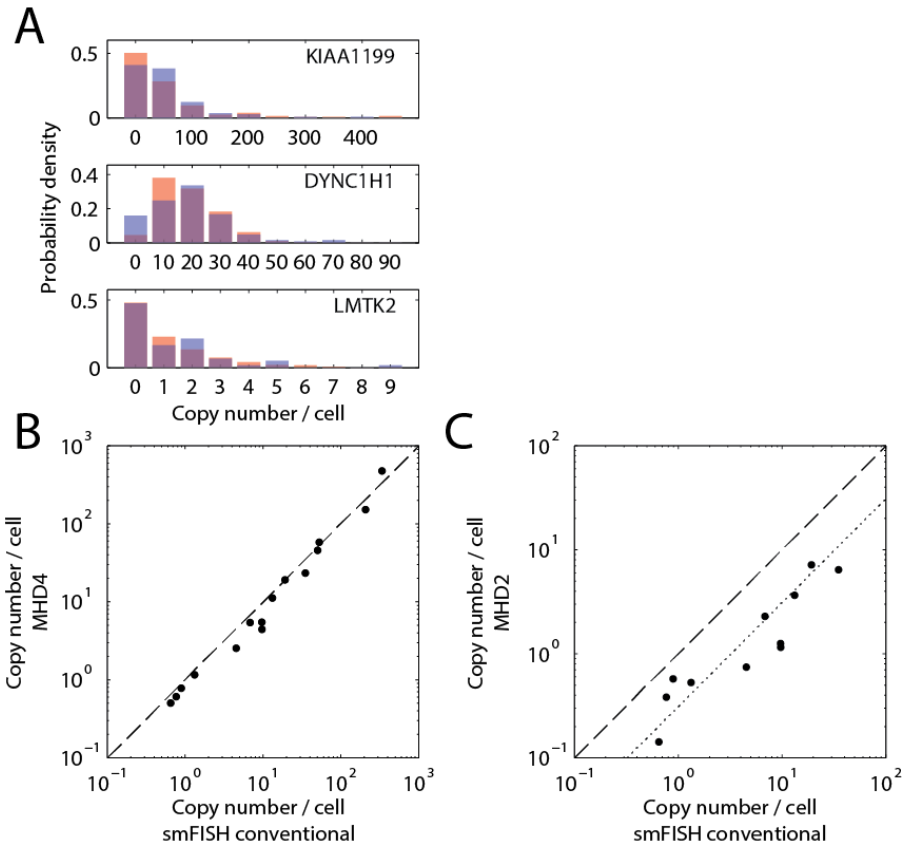

Fig. S8

Comparison of the MERFISH measurements with conventional smFISH results for a subset of genes. (**A**) The distributions of RNA copy numbers in single cells for three example genes KIAA1199, DYNC1H1, and LMTK2 in our high, medium and low abundance ranges, respectively. Red bars: distributions constructed from ~400 cells in the 140-gene measurements using the MHD4 codes. Blue bars: distributions constructed from ~100 cells in the conventional smFISH measurements. **(B)** Comparison of the average RNA copy numbers per cell measured in the 140-gene experiments using the MHD4 codes to those determined by conventional smFISH for 15 genes. The average ratio of the copy number measured using the MHD4 measurements to that measured using conventional smFISH is 0.82 ± 0.06 (mean ± SEM across 15 genes). The dashed line corresponds to the $y = x$ line. (**C**) Comparison of the average RNA copy numbers per cell measured in the 1001-gene experiments using the MHD2 code to those determined by conventional smFISH for 10 genes. The average ratio of the copy number measured using the MHD2 measurements to that measured using conventional smFISH is 0.30 ± 0.05 (mean ± SEM across 10 genes). The dashed line corresponds to the $y = x$ line and the dotted line corresponds to the $y = 0.30x$ line.

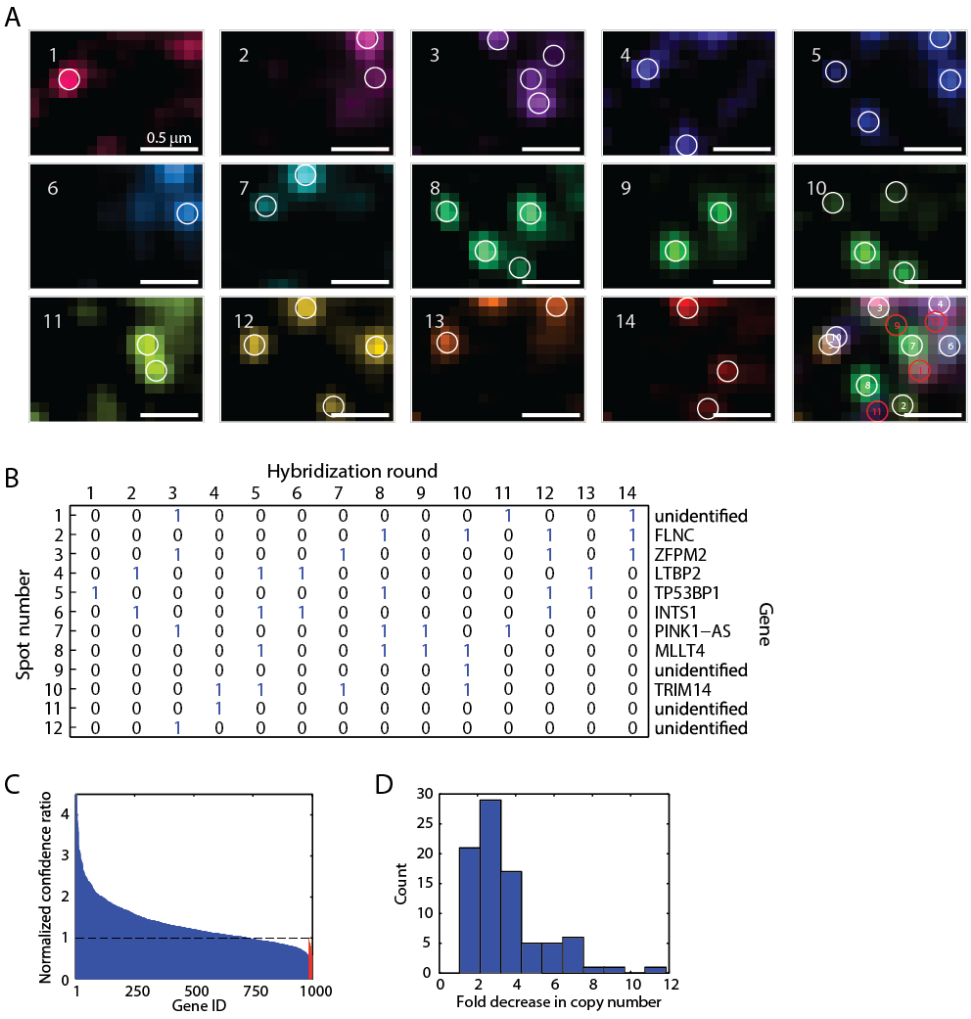

Fig. S9

Decoding and error assessment of the 1001-gene experiments. (**A**) Images of the boxed sub-region of the cell in Fig. 5A for each of 14 hybridization rounds. The final panel is a false colored, composite image of these 14 rounds. Circles indicate fluorescent spots that have been identified as potential RNA molecules. Red circles in the composite image indicate unidentifiable molecules, the binary words of which do not match any of the 14-bit MHD2 code words. (**B**) The corresponding binary word for each of the spots identified in (A) with the RNA species to which it is decoded. 'unidentified' implies that the measured binary word does not match any of the 1001 code words. (**C**) The normalized confidence ratios measured for the 985 RNA species (blue) and the 16 misidentification control words not targeted to any RNA (red). The normalized confidence ratio is defined as in Fig. 2F. (**D**) Histogram of the reduction in detected abundance of 107 genes present in both the 1001-gene experiments and the 140-gene experiments. "Fold decrease in copy number" is defined as the average number of RNA molecules per cell for each species measured in the 140-gene experiments divided by the corresponding average number measured in the 1001-gene experiments.

Table S1 (Provided as a separate Excel file)

Two different codebooks for the 140-gene experiments. The specific code words of the 16-bit MHD4 code assigned to each RNA species studied in the two shuffles of the 140-gene experiment. The "Genes" columns contain the name of the gene. The "Codewords" columns contain the specific binary word assigned to each gene.

Table S2 (Provided as a separate Excel file)

GO analysis of the gene groups with co-varying expression in 140-gene measurements. The 'Genes' column contains the names of the genes for each identified group. "Correlation_difference" is the difference in the average correlation coefficient between the specified gene and all others in that group and the average correlation coefficient between the specified gene and all others genes not in that group. "P_value_correlation" represents the significance (p-value) of this difference in average correlation coefficient as determined by a student's t-test. All of grouped genes have p-values that are substantially smaller than most of p-values of the ungrouped genes. "GO_ID" lists the GO IDs associated with the enriched GO terms for each group. "GO_terms" lists the names of these GO terms. "GO_enrichment" provides the measured enrichment, defined as the ratio of the fraction of genes within each group that have this term to the fraction of all measured genes that have this term. "P_value_GO" lists the significance (p-values) of the enrichment of these terms, calculated via the hypergeometric function. Only the top 10 statistically significantly enriched GO terms are listed for each groups. "Genes_with_little_or_no_annotation" lists genes in each group with little or no prior annotation for which we can now hypothesize function based on this grouping. Non-distinct GO terms such as "protein-binding" and "DNA-binding" are not reported as top GO terms even if they are the most statistically significant in the list. In order to better predict functions for transcription factors, general terms associated with transcription factor activity have also been excluded, such as "transcription, DNA-templated", "regulation of transcription, DNA-templated", and "sequence-specific DNA binding transcription factor activity".

Table S3 (Provided as a separate Excel file)

Codebook for the 1001-gene experiments. The specific code words of the 14-bit MHD2 code assigned to each RNA species studied in the 1001-gene experiments. The columns are defined as those listed in Table S1.

Table S4 (Provided as a separate Excel file)

GO analysis of the gene groups with co-varying expression in 1001-gene measurements. The columns are defined as those listed in Table S2. Here the p-values of the average correlation difference is greater than 0.05 for a small number of the grouped genes (marked red) suggesting that assignment to that group may not be statistically significant. The statistically most significantly enriched GO terms (maximum 10) are listed for each group. Several groups do not have statistically significant enrichment of GO terms and these groups are labeled green.

Table S5 (Provided as a separate Excel file)

Template sequences for the construction of all encoding probes. The “Experiment” column defines the experiment. The “Codebook” column lists the number of the codebook. The “Gene” column lists the name of the target genes. The “Index Primer 1” column lists the sequence of the first index primer. The “Common RT Primer” column lists the sequence of the common reverse transcription primer. This primer is not used in the 1001-gene experiments, and the index 1 primer is used instead as the reverse transcription primer. The “Readout 1” column lists the sequence of the first readout sequence. The “Targeting region” column lists the sequence of the region that targets cellular RNA. The “Readout 2” column lists the sequence of the second readout sequence. The “Index Primer 2” column lists the sequence of the second index primer. The sequences of the encoding probe templates are the concatenation of these sequences in order of the columns from left to right.